Soft Comput (2018) 22:1623–1639
https://doi.org/10.1007/s00500-016-2425-2

METHODOLOGIES AND APPLICATION

A novel anomaly detection algorithm for sensor data under uncertainty

Raihan Ul Islam[1] · **Mohammad Shahadat Hossain**[2] · **Karl Andersson**[1]

Published online: 9 November 2016

Abstract It is an era of Internet of Things, where various types of sensors, especially wireless, are widely used to collect huge amount of data to feed various systems such as surveillance, environmental monitoring, and disaster management. In these systems, wireless sensors are deployed to make decisions or to predict an event in a real-time basis. However, the accuracy of such decisions or predictions depends upon the reliability of the sensor data. Unfortunately, erroneous data are received from the sensors. Consequently, it hampers the appropriate operations of the mentioned systems, especially in making decisions and prediction. Therefore, the detection of anomaly that exists with the sensor data drew significant attention and hence, it needs to be filtered before feeding a system to increase its reliability in making decisions or prediction. There exists various sensor anomaly detection algorithms, but few of them are able to address the uncertain phenomenon, associated with the sensor data. If these uncertain phenomena cannot be addressed by the algorithms, the filtered data into the system will not be able to increase the reliability of the decision-making process. These uncertainties may be due to the incompleteness, ignorance, vagueness, imprecision and ambiguity. Therefore, in this paper we propose a new belief-rule-based association rule (BRBAR) with the ability to handle the various types of uncertainties as mentioned.The reliability of this novel algorithm has been compared with other existing anomaly detection algorithms such as Gaussian, binary association rule and fuzzy association rule by using sensor data from various domains such as rainfall, temperature and cancer cell data. Receiver operating characteristic curves are used for comparing the performance of our proposed BRBAR with the aforementioned algorithms. The comparisons demonstrate that BRBAR is more accurate and reliable in detecting anomalies from sensor data under uncertainty. Hence, the use of such algorithm to feed the decision-making systems could be beneficial. Therefore, we have used this algorithm to feed appropriate sensor data to our recently developed belief-rule-based expert system to predict flooding in an area. Consequently, the reliability and the accuracy of the flood prediction system increase significantly. Such novel algorithm (BRBAR) can be used in other areas of applications.

Keywords Internet of Things · Wireless sensor networks · Anomaly detection · Flood prediction · Belief-rule-based expert systems

Communicated by V. Loia.

✉ Karl Andersson
karl.andersson@ltu.se

Raihan Ul Islam
raihan.ul.islam@ltu.se

Mohammad Shahadat Hossain
hossain_ms@cu.ac.bd

[1] Pervasive and Mobile Computing Laboratory, Luleå University of Technology, 931 87 Skellefteå, Sweden

[2] Department of Computer Science and Engineering, University of Chittagong, Chittagong 4331, Bangladesh

1 Introduction

Nowadays, wireless sensors are deployed in large scale to monitor various environmental parameters such as rainfall, water level, humidity, soil moisture and temperature (Ahmad et al. 2013; Zhang et al. 2011). The collected sensor data can be used in various expert systems to support decision-making processes or to predict the occurrence of an event such as flooding (Fang et al. 2014). The wireless sensors are

considered due to their low power consumption, low cost and protocol standardization (Palattella et al. 2013; Ahmad et al. 2013; Seal et al. 2012; Khedo 2013; Atzori et al. 2010). Usually, such expert systems are helpful where the events under investigation change rapidly and their prediction cannot be made in advance. Flooding can be considered as an example of such event, which has the highest capability to bring sufferings to the human beings and therefore, its assessment of risk is very important (Hossain and Davies 2001, 2004, 2006; Vladimirova and Yuhaniz 2011; Gnecco et al. 2016). Hence, wireless sensor network technologies have been used to collect flood-related data, and eventually, they are fed into Decision Support Systems (DSSs) to generate different decision scenarios and to predict flooding in an area (Andersson and Hossain 2014, 2015; González et al. 2013; Demeritt et al. 2013).

However, the accurate and appropriate risk scenario generations by these systems (Aziz and Aziz 2011; Adefisan et al. 2015) as well as the flood prediction are found to be not reliable due to the erroneous and misleading nature of sensor data (Pappenberger et al. 2006). The reason for this is that sensor data may contain missing data, duplicated data or inconsistent data due to the resource constraints such as battery power (Sheltami et al. 2016; Xu et al. 2015), computational and memory capacities (Bajaber and Awan 2010) as well as communication bandwidth (Thombre et al. 2016). Hence, the data generated by the sensor nodes become unreliable and inaccurate. In addition to this, in harsh environment where sensors are deployed in unprotected way, causing malfunction and this may result in noisy, missing and redundant data (Chen et al. 2006). Moreover, the sensors are vulnerable to malicious attacks such as denial of service attacks, black hole attacks and eavesdropping (Perrig et al. 2004; Langin and Rahimi 2010; Fiore et al. 2013).

The presence of missing value, duplicate or inconsistency with the sensor data leads to the creation of different types of uncertainty such as incompleteness, ignorance, vagueness, imprecision and ambiguity. The resource constraints of sensors cause some data to be missed, causing ignorance and ambiguity. The malfunction causes the sensor data to be incomplete. Moreover, vagueness is caused in sensor data by inaccuracy due to malicious attack, and imprecision is caused by less precise data reading from sensor due to lack of battery power (Rajasegarar et al. 2008). The presence of uncertainty with the sensor data resulting from the factors mentioned above may cause anomaly in the sensor data. Hence, the data become unreliable, and if they are not filtered before feeding to the expert systems, the results generated from such systems may become inaccurate. Therefore, it is necessary to use appropriate techniques to handle anomalous data with the capability of handling different types of uncertainty in an integrated framework.

Therefore, it is necessary to ensure the reliability and accuracy of the sensor data before using it in any expert system. By using anomaly detection techniques, we can ensure reliability and accuracy of the data. According to Gnecco et al. (2016), anomalies are patterns in data that do not conform to a well-defined notion of normal behaviour. There are different techniques of anomaly detection, based on the model used; these are parametric (statistical) and nonparametric model-based anomaly detection techniques (Chandola et al. 2009). In the parameter techniques, data are analysed using density distribution and which data have low relevance with the distribution are considered as anomalies. Multivariate Gaussian method is an example of statistical model-based anomaly detection technique. Statistical model works well when distribution of the data is known, which is rare for sensor data. Rule-based techniques are the examples of nonparametric approaches (Chandola et al. 2009). In the rule-based techniques, rules are generated based on the data. Each of the rules is given a weight value-based on the frequency of the rule in data, and anomalous data are detected using some threshold values. Association Rule mining (He et al. 2004) and Fuzzy Rule Base Association Rule mining (Weng 2011) are the examples of rule-based techniques. However, these rules do not take into account of the uncertainty phenomena of the sensor data.

However, Gaussian method, is a statistical-based approach, is unable to handle uncertainty due to randomness, ignorance as well as fuzziness. On the other hand, rule-based approach such as association rule uses assertive knowledge, which can be evaluated either true or false. Hence, this approach is unable to address uncertainty due to fuzziness, ignorance or incompleteness. Fuzzy logic can handle uncertainty due to fuzziness but unable to handle ignorance and incompleteness. It is also unable to handle uncertainty due to ignorance in fuzziness. Hence, none of the mentioned methods can handle all types of uncertainty in an integrated framework. Hence, in this research we are proposing a novel belief-rule-based anomaly algorithm with the capability of handling the mentioned uncertainties in an integrated framework. Eventually, this will accurately detect the anomalous data and filtered to feed the DSSs to predict an event accurately and also its risks.

The remainder of this paper is structured as follows: Sect. 2 surveys related work on anomaly detection, while Sect. 3 provides the overview of the new anomaly detection technique named, Belief Rule-Based Association Rule (BRBAR). Section 4 presents the integration of filtered sensor data into belief-rule-based expert system. Section 5 reports the experimental results and evaluation of BRBAR, while Sect. 6 concludes the paper.

2 Related work

Research on anomaly detection has been going on for a long time, specifically in the area of statistics (Chandola et al. 2009). Multivariate Gaussian, a statistical-based anomaly detection algorithm was proposed by Barnett and Lewis (1994), Barnet (1976), and Beckman and Cook (1983). The underlying principal of this method is that the anomalous data should be detected by using a parametric or Gaussian distribution as well as by using probability density function. In this method, the latter is used to calculate the anomaly score of the data. A threshold value is then used to determine anomalous data from the anomaly score. In Gaussian-based anomaly detection technique, it is assumed that the dataset will follow the Gaussian distribution. Dataset, which are unimodal, symmetric, asymptotic in nature usually provides normal distribution. If the dataset cannot fully follow the distribution, then the inaccuracy in anomaly detection may be noticed. This inaccuracy causes uncertainty in anomaly detection. Therefore, statistical-based anomaly detection algorithms such as Gaussian distribution, fails to take in to account of uncertainty. In addition, all datasets cannot be modelled using Gaussian distribution if the data points are not clustered around the mean value of the dataset. Furthermore, threshold parameter might be difficult to determine as the difference between nonanomalous and anomalous data might be very close (Patcha and Park 2007). Moreover, if the dataset is asymmetric and bimodal, then the proper detection of anomalous data is difficult to obtain using Gaussian distribution. However, the nature of the sensor data is asymmetric or bimodal. Therefore, statistical-based anomaly detection approach will not be efficient for anomaly detection. Alternatively, knowledge-based approach, based on the frequency of the data points in the datasets, provides better detection of anomalous data. Since the sensor data are asymmetric in nature, the determination of the frequency of data can be used to develop rules. This in turn could form knowledge base and thus, can be used to detect anomalies in sensor data by using various knowledge-based approaches. Therefore, in the following section knowledge-based approaches will be investigated to demonstrate their strength to detect anomalies in sensor data.

Generally, in rule-based or knowledge-based anomaly detection, the anomaly detector uses predefined rules to classify data points as anomalies or normal data. There exist various types of rule-based approaches such as association rule, fuzzy association rule to detect anomaly in the sensor data (Chandola et al. 2009).

Association rule is a rule-based approach for data mining. It was first proposed by Agrawal and Srikant (1994) to detect frequent item sets from database of items in a shop purchased by people. Association rule is expressed as a form of $X \rightarrow Y$, where X, Y are subsets of items. The rule implies that if a person purchases X item sets, then the person might also purchases Y item sets. Using the above-mentioned algorithm at first, frequent itemsets were detected using minimum support and then from the frequent itemsets using minimum confidence association rules are discovered. However, during finding the frequent itemsets crisp values are considered, which lack the capability of addressing the issues of different types of uncertainties like ignorance, incompleteness, ambiguity, vagueness and imprecision.

An association rule-based anomaly detection technique is proposed in He et al. (2004). The authors present a new method to detect anomaly by discovering frequent patterns from the dataset. In this method, each data point in the dataset is considered as a transaction. Therefore, the transactions that contain less frequent patterns are detected as anomaly. This method defines a measure, called FPOF (Frequent Pattern Outlier Factor), to detect the anomalous transactions. However, the method can well handle precise data, and hence, it is not well suited where the nature of the data contains fuzziness. In addition, sensor data contain various types of uncertainty such as ignorance, incompleteness, ambiguity, vagueness and imprecision for the reasons as explained in the previous section. Thus, by using this method the appropriate rules cannot be mined, and hence, the detection of the anomaly exits in the sensor data.

Sensor data can be viewed as a large volume of real-valued data collected from sensor nodes. The characteristics of these data depend on the attributes of data as well as on the correlation between the data in space and time. Each sensor node might have one or more sensors. A sensor node with one temperature sensor, which can be considered as providing univariate-attributed data. On the other hand, a sensor node consists of temperature and humidity sensor, and these can be considered as multivariate attributed data. It is comparatively easier to detect outlier from univariate-attributed data as one need to consider one type of data. However, for multivariate attributed data to detect anomaly multiple types of data need to be considered together. Moreover, special and temporal correlation with the collected data also influences in anomaly detection in sensor data. Temporal correlation implies the reading of sensor data in one instant is related to the previous instant of the time. On the contrary, special correlation implies that a correlation exists among the data gathered from geographically closely deployed sensors (Zhang et al. 2010).

There exist various techniques (Weng 2011; Rajeswari et al. 2014; Ruiz et al. 2014; Muyeba et al. 2008) to detect anomaly in sensor data by using Fuzzy association rules. In fuzzy association rule, the data points are converted to fuzzy values using membership function. Fuzzy association rules are then generated based on frequent data points or rare frequent data. Using the generated rules, anomalous data are detected from the sensor data. Fuzzy sets overcome the problem of overestimate or underestimate the boundary val-

ues by using membership function. Fuzzy logic is capable of handling uncertainty due to imprecision, ambiguity and vagueness but not the others.

Weng (2011), proposes an anomaly detection technique based on rare data pattern instead of frequent data pattern. This methodology is able to discover more interesting and valuable patterns from the data and then the association rule-based technique. However, experts assign membership function and four parameters (e.g. minimum support, maximum support, maximum rank and minimum confidence) in this study. This makes the system more human dependent. Moreover, the proposed algorithm is not able to address ignorance and incompleteness due to the limitation of fuzzy logic.

Rajeswari et al. (2014) studied anomaly detection on educational data using fuzzy association rule mining. The authors argue that fuzzy logic handles data better and it can calculate dynamically the four parameters rather than using predefined values, mentioned above, produces better results. By using a modified Fuzzy Apriori Rare Itemsets Mining (FARIM) (Weng 2011) algorithm, teachers can more easily detect weak students and give them extra coaching. The proposed method also did not address the ignorance and incompleteness.

Ruiz et al. (2016) introduced the notion of fuzzy exception and fuzzy anomalous rule for recognition of various types of deviations often associated with the common patterns which usually are hidden in data affected by some fuzziness. A new approach for mining such rules is presented, whereas important advantages include obtaining more understandable results and that the mining process can be parallelized. The authors present an algorithm along with experiments performed in data where some numerical attributes were fuzzified. The authors concluded that the proposed fuzzy rules give some insights on the exception and anomaly detection in credit payments.

Martí et al. (2015) proposed an anomaly detection algorithm based on sensor data for petroleum industry applications. They have dealt with the problem of detecting anomalies in turbo machines used in offshore oil platforms. The algorithm is composed of a novel segmentation algorithm, which is named YASA, and one-class support vector machine (SVM). The authors have compared YASA with one-class SVM and the approach currently used by their industry partners. The results show that the combination of YASA and one-class SVM was able to outperform the other approaches. However, the proposed algorithm lacks addressing different types of uncertainty associated with sensor data, such as incompleteness, ignorance, vagueness, imprecision and ambiguity.

In summary, Gaussian-based anomaly detection algorithm provides a mechanism for detecting anomaly from multivariate sensor data without any prior knowledge of the data. The algorithm assumes that the sensor data follow normal or Gaussian distribution. However, this is not true for every sensor data, in that case the Gaussian-based anomaly detection algorithm does not detect anomalous data efficiently. Moreover, the algorithm does not have any mechanism to detect and address the uncertainty due to ignorance, incompleteness, ambiguity, vagueness and imprecision. Association rule-based anomaly detection does not have dependency on the data distribution. However, it also lacks on addressing uncertainty. Fuzzy-based association rule provides mechanism to address the problem of overestimate or underestimate the boundary values by using membership functions. These techniques are capable of handling uncertainty due to imprecision, ambiguity and vagueness but not the others by using fuzzy set. Therefore, a novel algorithm is required to address all types of uncertainty that exist with the sensor data by using an integrated framework to detect anomalous data. Hence, the following section describes a novel BRBAR with the ability to handle various types of uncertainty like ignorance, incompleteness, ambiguity, vagueness and imprecision for detecting anomalous sensor data.

3 An overview of belief-rule-based association rule

In this section, binary association rule as well as fuzzy association rule to detect anomaly in sensor data will be introduced. The limitations of these approaches in handling various types of uncertainties will be demonstrated. Then, BRBAR will be introduced which has the capability to handle all types of uncertainty in an integrated framework.

3.1 Binary association rule

Binary association rule is created from frequent itemsets of transactions occurring in a database. Itemsets is a collection of items available in the database. There are two main parameters, namely support and confidence. Support can be defined as the frequency of itemsets in whole database divided by number of transactions (Agrawal and Srikant 1994). Confidence can be defined as the frequency of itemsets in the rule divided by the frequency of itemsets in antecedent part of the rule (Agrawal and Srikant 1994). Support is like finding the probability of an itemsets in the database, and confidence is the conditional probability (Rajeswari et al. 2014).

Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of m literals called *items* and the database $D = \{t_1, t_2, \ldots, t_n\}$ a set of n transactions, each consisting of a set of items from I. An itemset X is a nonempty subset of I. The length of itemset X is the number of items in X. An itemset of length k is called a *k-itemset*. A transaction $t \in D$ is said to contain itemset X if $X \subseteq t$. The *support* of itemset X is defined as $support(X) = ||t \in D | X \subseteq t||/||t \in D||$.

The support of association rule is shown in expression (1).

$$support(A \implies B) = support(A \cup B) \tag{1}$$

The confidence of association rule is shown in expression (2).

$$conf(A \implies B) = support(A \cup B)/support(A) \tag{2}$$

The binary association rule is shown in expression (3).

$$A \implies B \tag{3}$$

Here, A and B indicate itemsets. A and B represent the antecedent and consequent part of an association rule, respectively.

He et al. (2004) proposed anomaly detection technique based on frequent itemsets. Frequent itemsets discovered by association rule algorithm provide common pattern of the dataset. The infrequent itemsets intuitively refer to anomalies. They propose a measure called *FPOF (Frequent Pattern Outlier Factor)* to detect the anomaly, which is shown in expression (4) (He et al. 2004).

$$FPOF(t) = \frac{\sum_{x} support(X)}{||FPS(D, minisupport)||} \tag{4}$$

where $X \subseteq t$ and $X \in FPS\ (D, minisupport)$

Here, all frequent patterns are denoted as: *FPS (D, minisupport)*.

In binary association rule, the support is calculated by computing the frequency of items. An association rule, as shown in expression (3), is evaluated as true or false and hence, does not provide scope of considering any types of uncertainty. However, sensor data contain different types of uncertainty, and thus, association rule is not appropriate to detect anomalies in sensor data. In addition, finding association rules from sensor data with quantitative attributes are problematic due to the poor semantic content to define the sensor data which creates vagueness and ambiguity (He et al. 2004). Moreover, the binary association rules are sensitive to small value changes which is a regular phenomena in sensor data. Association rule also has a tendency to overestimate or underestimate the boundary value (Rajeswari et al. 2014) during the process of transforming the transaction database to a binary database by partitioning the attribute values (Chen and Chen 2007). Above problems can be address by fuzzy association rule, which will be shown in next section. Furthermore, mining association rules are computationally costly (Wijsen and Meersman 1998) as large number of binary association rules are generated during binary association rule mining process.

3.2 Fuzzy association rule

Fuzzy association rules are created from quantitative data, in which each quantitative item is transformed into fuzzy set and fuzzy operations are used to find fuzzy association rules.

A fuzzy association rule is represented as shown below.

$$(x_i\ is\ a_1) AND (x_i\ is\ a_2) \implies (y_i\ is\ m_k) \tag{5}$$

Here, x and y stand for antecedent and consequent attributes. a and m represent the referential values.

The quantitative values form sensor data are represented using linguistic labels or referential values by fuzzy sets in the process of mining fuzzy association rules from sensor data. For example, the values of a sensor data attribute like temperature might be represented using different linguistic labels such as very high, high, medium and low in fuzzy set. This helps to represent the semantic content of the sensor data more efficiently than the binary association rule by providing meaningful linguistic labels of sensor data (Chen and Chen 2007). Moreover, using fuzzy membership functions of fuzzy sets overestimation or underestimation of the boundary values of binary association rule can be addressed by allowing partial membership to different fuzzy sets (Dhanya and Kumar 2009).

Weng (2011) proposed an anomaly detection technique using fuzzy set. The proposed technique finds anomalous data using rare itemsets instead of frequent itemsets. The anomaly detection technique uses a function named rank to find rare itemsets from the transaction dataset. Let us consider a database D consists of a set of transaction A_{tid} of sensor data. A fuzzy rule of itemset is denoted as B. The rank $Rank_D(B)$ of the fuzzy rule B in database D can be defined as follows.

$$Rank_D(B) = \frac{\sum_{tid=0}^{|DB|} rnk(A_{tid}, B)}{|DB|} \tag{6}$$

Here, DB is the subset of transactions covered by the fuzzy rule B in the database D.

The support is calculated by using following expression

$$sup_D(B) = \frac{\sum_{tid=0}^{|D|} sup(A_{tid}, B)}{|D|} \tag{7}$$

where $|D|$ is the total number of transactions in database D. According to Hossain et al. (2014), fuzzy set addresses three types of uncertainty due to vagueness, ambiguity and imprecision using membership function and referential values. However, the consequent part of the fuzzy association rule,

as shown in expression (5), considers only one attribute at time. Therefore, it is not able to address all types of uncertainty in sensor data, and also this leads to generation of higher number of rules. The inference mechanism of the fuzzy association rule does not have any option of determining the uncertainty in sensor data like incompleteness. However, the sensor fails to send data due to network resource constrain or malicious attack and thus, causing uncertainty like incompleteness. Therefore, fuzzy association rule is not fully suitable for anomaly detection. Most of the problems mentioned above will be addressed by the new belief-rule-based association rule described in the next section.

3.3 Belief-rule-based association rule

The belief rule base (BRB) is an extension of traditional IF-THEN rule base. A belief rule has antecedent part and consequent part. Antecedent attribute takes referential values, and possible belief degrees are associated with the consequent of a belief rule. The rule weight, antecedent attribute weight and belief degrees are knowledge representation parameters used in BRB to capture the uncertainty.

A belief rule can be defined as:

$$R_k: \begin{array}{l}\text{IF Rainfall is Medium AND Rainfall Duration is High THEN Meteorological Condition is}\\ \{(\text{Severe}, 0.0), (\text{Moderate}, 0.4), (\text{Low}, 0.6)\}\end{array} \quad (8)$$

In the above rule, Rainfall and Rainfall Duration are the antecedent attributes, while Medium and High are the referential values. Meteorological Condition is the consequent attribute with referential values such as severe, moderate and low. This rule is complete because the summation of degree of belief associated with each referential value of the consequent attribute is one. If the summation is less than one, then the rule is considered as incomplete, which may be due to incomplete information or ignorance. The relationship between antecedent attributes and the consequent attribute is nonlinear, which is linear in case of IF-THEN rule. Moreover, in general the sensor data that are gathered from the environment are nonlinear in nature (Xiea et al. 2014; Islam et al. 2015). Therefore, belief rules can efficiently be used to represent the sensor data.

Inference mechanism is utilised to generate belief rules from sensor data. The inference procedures consist of various steps including input transformation, rule activation weight calculation, belief update and rule aggregation using evidential reasoning approach (Hossain et al. 2014, 2015c; Rahaman and Hossain 2013). The task of input transformation consists of distributing the input data over the referential values of the attribute of a rule, which is called matching degree. Once the matching degree is assigned, the rules are called packet antecedent, and they become active and reside in the short-term memory while the rule base resides in the long-term memory. The total degree or the combined matching degree α_k, to which the input matches the whole antecedent part of kth rule, can be calculated by using the following expression (Hossain et al. 2015b).

$$\alpha_k = aggr((\delta_{k1}, \alpha_1^k), \ldots, (\delta_{kT_k}, \alpha_{T_k}^k)) \quad (9)$$

where $aggr$ is an aggregation function which should be selected carefully. Following simple weighted multiplicative aggregation function can be used as an aggregation function (Hossain et al. 2015b).

$$\alpha_k = \prod_{i=1}^{T_k} (\alpha_i^k)^{\bar{\delta}_{ki}} \quad (10)$$

where $\bar{\delta}_{ki} = \frac{\delta_{ki}}{max_{i=1,\ldots,T_k^{\{\delta_{ki}\}}}}$ so that $0 \leq \bar{\delta}_{ki} \leq 1$

Here, T_k is the total number of antecedent attributes in the kth rule. The activation weight w_k for kth rule can be generated by the following expression.

$$w_k = \frac{\theta_k \alpha_k}{\sum_{i=1}^{L} (\theta_i \alpha_i)} \quad (11)$$

Here, θ_k represents the rule weight, and α_k represents the combined matching degree of the kth rule.

It is interesting to note that each rule does not have the same weight in calculating the referential values of the consequent attribute. This activation weight will be zero if the kth rule is not activated.

When an input data for any of the antecedent are ignored or missing, then the belief degree associated with each rule in the rule base should be updated. Therefore, in belief update procedure the belief degree of each of the rule is updated using following expression (Hossain et al. 2015b).

$$\beta_{ik} = \bar{\beta}_{ik} \frac{\sum_{t=1}^{T_k} (\lambda(t,k) \sum_{j=1}^{J_t} (\alpha_{tj}))}{\sum_{t=1}^{T_k} \lambda(t,k)} \quad (12)$$

where

$$\lambda(t,k) = \begin{cases} 1 & \text{if } t\text{th attribute is used in defining rule } R_k (t = 1, \ldots, T_k) \\ 0 & \text{otherwise} \end{cases}$$

Here, $\bar{\beta}_{ik}$ represents the original belief degree, while the updated belief degree is β_{ik} of kth rule. α_{tj} represents the degree to which the input value belongs to an attribute.

Furthermore, the aggregation of the rules is carried out by using either analytical or recursive evidential reasoning algorithm (Yang et al. 2006; Xu et al. 2007). It is preferable to use analytical approach instead of recursive approach since it is computationally efficient (Yuan et al. 2002; Yang and Sen 1994). Using the analytical ER algorithm (Wang et al. 2006), the final belief degree β_j is calculated using following expression.

$$\beta_j = \frac{\mu \times \left[\prod_{k=1}^{L}\left(\omega_k \beta_{jk} + 1 - \omega_k \sum_{j=1}^{N} \beta_{jk}\right) - \prod_{k=1}^{L}\left(1 - \omega_k \sum_{j=1}^{N} \beta_{jk}\right)\right]}{1 - \mu \times \left[\prod_{k=1}^{L} 1 - \omega_k\right]} \tag{13}$$

where

$$\mu = \left[\sum_{j=1}^{N}\prod_{k=1}^{L}\left(\omega_k \beta_{jk} + 1 - \omega_k \sum_{j=1}^{N} \beta_{jk}\right) - (N-1) \times \prod_{k=1}^{L}\left(1 - \omega_k \sum_{j=1}^{N} \beta_{jk}\right)\right]^{-1}$$

Here, ω_k represents the activation weight of the kth rule, whereas the belief degree associated with one of the consequent reference values is denoted by β_j.

The final values can be converted into crisp values by using the utility score associated with each referential value to obtain the final result. Hence, by summing the belief degrees of the referential values of the consequent part of the expressions (8) should be one if all the sensor data for the antecedent part are available which address the uncertainty due to incompleteness. The expression (12) addresses the uncertainty due to ignorance or missing values from sensors by updating the belief degree of each of the rules during belief update procedure. Moreover, the uncertainty due to vagueness, imprecision and ambiguity is addressed by the expression (13) during the process of rule aggregation (Wang et al. 2006). As we discussed in the previous sections, sensor data contain anomalous data due to different kinds of uncertainty like incompleteness, ignorance, vagueness, imprecision and ambiguity. From the above discussion, it can be argued that belief-based rule and inference mechanism addresses all type of uncertainty.

However, the above inference procedures, which consist of input transformation, rule activation weight calculation, belief update and rule aggregation of belief rule base cannot be directly applied to discover belief rules from sensors data. The reason for this is that it is not necessary to have initial rule base in case of sensor data, because the objectives of sensor data mining are to discover the sets of belief rules

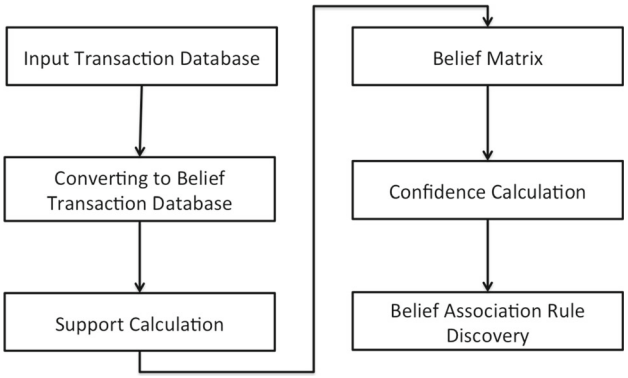

Fig. 1 Flow chart of belief association rule discovery

which in turn will act as the initial belief rules to represent the knowledge base of an expert system. Hence, it is necessary to investigate appropriate inference methods. However, in the light of belief-rule-based inference procedures to discover initial belief rules, the task of input transformation can be carried out by developing input transaction database as well as by converting the transaction database into belief transaction database. Since the calculation of support, as discussed both in case of binary and fuzzy rules [(6), (1)], it is necessary to develop a procedures to calculate support for belief transaction database. In addition, it is also necessary to calculate the confidence of the belief transaction database, which can be achieved by developing belief matrix and hamming distance calculation. This will allow the calculation of confidence of each transaction of belief database. Eventually, belief association rule could be discovered for sensor data, which act as the initial belief rule base for an expert system. It can be demonstrated that by using the belief association rules, and the confidence values anomalies from sensor data can be removed. The above procedures diagrammatically demonstrated in Fig. 1, and we would like to define whole procedures as the BRBAR. Each of the procedures as shown in the Fig. 1 will be discussed in detail.

As a first step, the sensor data among which we want to find anomaly, each data points are given an id. Henceforth, each transaction is entered into the transaction database. The transaction database is then converted into belief transaction database by input transformation (Andersson and Hossain 2015). Support of the sensor data is calculated in the next step named support calculation. Subsequently, a belief matrix is created. Hamming distance is then calculated to find the differences among the transactions. Confidence of each of the transaction is calculated. Using the belief matrix as well as the confidence value belief-rule-based association rules are discovered which are free from any anomalous sensor data and thus can be use as initial belief rules in a BRB (Belief Rule Base). This demonstrates a novel way of extracting belief rules from the sensor data. In addition, to support the mining of sensor data, it is necessary to develop a novel way

Table 1 Sample of transaction database

Transaction ID	Rainfall	Temperature
t1	7	27
t2	1900	32
t3	450	40
t4	290	32
t5	190	30
t6	510	30
t7	571	27
t8	349	31
t9	259	20
t10	85	24

to calculate support and confidence as discussed while presenting binary and fuzzy association rules. However, these techniques of calculating both support and confidence cannot consider different types of uncertainty. Consequently, it is necessary to develop novel methods by incorporating of different types of uncertainty in calculating support and confidence. Hence, this research demonstrates novel methods of calculating support and confidence by incorporating different types of uncertainty as will be demonstrated below.

3.3.1 Input transaction database

Input transaction database will contain all sensor data. To identify the sensor data, each row of the data is given a unique identification number, named as transaction ID. However, the sensor data are quantitative in nature. Therefore, these data contain uncertainty like imprecision, vagueness. Moreover, these data are also semantically poor. Therefore, to address the above uncertainties and to address poor semantic content referential values and linguistic labels are introduced. The data from input transaction database will be used to support to get belief transaction database, which will contain sensor data with referential values. For simplicity, a sample transaction database is presented in Table 1 which contains ten rows of data. The database has three attributes which are shown in Table 1. These are transaction ID, rainfall and temperature. From this database, anomalous data of rainfall and temperature, which are collected by sensors, will be discovered. Linguistic labels and referential values are defined for rainfall and temperature to address the issue of poor semantic content, and this will remove the above-mentioned uncertainties. Tables 2 and 3 provide an example of linguistic labels and referential values derived by discussing with experts.

3.3.2 Converting to belief transaction database

Belief transaction database can be defined as the collection of referential values of the sensor data. Data from transactional database are taken as input, and then the sensor data are converted into referential values using utility function (Andersson and Hossain 2015; Wang et al. 2006; Hossain et al. 2015c). This facilitates the computational procedure of support calculation of BRBAR. This step allows to address uncertainty due to ambiguity, vagueness and imprecision of sensor data by distributing the degree of belief into the referential values. Converting input transaction database into belief transaction database resembles the input transformation of inference mechanism of belief rule base. Subsequently, the referential values are used for calculating support values of the sensor data, which is an essential step for mining anomalies. However, the details of this step will be presented in the next section. The expression (14) and (15) are used as utility function. In the expressions (14) and (15), x_i represents the ith referential value of an attribute, x_{i+1} represents the $(i+1)$th referential value, and a represents the sensor data

Table 2 Labels and referential values for rainfall

Rainfall					
Labels	No Rainfall	Low	Medium	High	Very High
Referential value	0	500	1000	1500	2000

Table 3 Labels and referential values for temperature

Temperature					
Labels	Very low	Low	Medium	Hot	Very Hot
Referential value	0	10	20	30	40

$$TransformValue(x_{i+1}) = \frac{|x_i - a|}{|x_i - x_{i+1}|} \tag{14}$$

$$TransformValue(x_i) = 1 - TransformValue(x_{i+1}) \tag{15}$$

where $x_i < a < x_{i+1}$

$$\left(\beta_{jk} \geq 0, \sum_{j=1}^{N} \beta \leq 1\right) \tag{16}$$

An example of the belief transaction database for rainfall and temperature is shown in Tables 4 and 5, respectively. The Column 1 of the Tables 4 and 5 refer to the transaction ID. Attribute name and the sensor value are shown in Column 2 of the Tables 4 and 5. Column 3 to 7 of the same table shows the referential values for the attributes. Row 1 of Table 4 shows that the transaction ID t_1 and the attribute a_1 (rainfall) have value 7. By using expression (14) and (15) for the sensor data 7 the degree of belief associated with referential values that can be obtained is {No rainfall (0.986), low

Table 4 Belief transaction database for rainfall

Transaction ID	Attribute:Value	No Rainfall	Low	Medium	High	Very High
t1	a1:7	0.986	0.014	0.0	0.0	0.0
t2	a1:1900	0.0	0.0	0.0	0.2	0.8
t3	a1:450	0.1	0.9	0.0	0.0	0.0
t4	a1:290	0.42	0.58	0.0	0.0	0.0
t5	a1:190	0.62	0.38	0.0	0.0	0.0
t6	a1:510	0.0	0.98	0.02	0.0	0.0
t7	a1:571	0.0	0.858	0.142	0.0	0.0
t8	a1:349	0.302	0.698	0.0	0.0	0.0
t9	a1:259	0.482	0.518	0.0	0.0	0.0
t10	a1:85	0.83	0.17	0.0	0.0	0.0

Table 5 Belief transaction database for temperature

Transaction ID	Attribute:Value	Very low	Low	Medium	Hot	Very Hot
t1	a2:27	0.0	0.0	0.3	0.7	0.0
t1	a2:32	0.0	0.0	0.0	0.8	0.2
t3	a2:40	0.0	0.0	0.0	0.0	1.0
t4	a2:32	0.0	0.0	0.0	0.8	0.2
t5	a2:30	0.0	0.0	0.0	1.0	0.0
t6	a2:30	0.0	0.0	0.0	1.0	0.0
t7	a2:27	0.0	0.0	0.3	0.7	0.0
t8	a2:31	0.0	0.0	0.0	0.9	0.1
t9	a2:30	0.0	0.0	0.0	1.0	0.0
t10	a2:85	0.0	0.0	0.6	0.4	0.0

(0.014), medium (0), high (0), very high (0)}. Moreover, the summation of degree of belief associated with the referential values is equal to one, which shows completeness according to expression (16).

3.3.3 Support calculation

Support calculation of BRBAR is defined as a function of sensor data and referential values in respect of belief transaction database. Sensor data and referential values are taken as input for support calculation, and the frequency of the sensor data with respect to the belief transaction database is provided. Binary [expression (1)] and fuzzy association rule [expression (7)] based anomaly detection algorithms use support function to find the probability of an itemset in the database. In the case of BRBAR, referential values of sensor data are also included in support calculation. Consequently, the support of BRBAR has the ability to address uncertainties like incompleteness, ignorance, vagueness, imprecision and ambiguity. In addition, fuzzy association rule considers only one of the referential values of consequent part in a rule, and hence, it is not able to address uncertainty like ignorance and incompleteness (Hossain et al. 2015b). On the contrary, belief association rules consider referential values of consequent attribute embedded with degree of beliefs as shown in Tables 4 and 5. The inclusion of this phenomenon with belief association rules provides strength to address the issues of ignorance and incompleteness of sensor data.

The support calculation for BRBAR is shown in expression (17). This helps to address the uncertainty of sensor data like imprecision, ambiguity and vagueness, because the expression (17) uses the referential values. In the expression (17), x_i represents the value of the sensor data. ref_val refers to referential values for the x_i.

$$Support(x_i) = \frac{\sum_{i,j}^{n} x_i * ref\_val_j}{|No\ of\ Transactions|} \tag{17}$$

Support of rainfall and temperature data collected by sensors is shown in Table 6, which is calculated using the expression (17). Column 1 of the Table 6 shows rainfall data in Sub-Column 1 named value and the support values on the Sub-Column 2 named support. Consecutively, temperature data and support values are shown on the second column of the Table 6. As for example, support calculation for rainfall value 7 is $((7 \times 0.986) + (7 \times 0.014) + (7 \times 0) + (7 \times 0) + (7 \times 0))/10 = 0.7$.

Table 6 Support calculation for rainfall and temperature

Rainfall		Temperature	
Value	Support	Value	Support
7	0.7	27	2.7
1900	190.0	32	3.2
450	45.0	40	4.0
290	29.0	32	3.2
190	19.0	30	3.0
510	51.0	30	3.0
571	57.1	27	2.7
349	34.9	31	3.1
259	25.9	30	3.0
85	8.5	85	2.4

3.3.4 Creating belief matrix

Belief matrix can be defined as the combination of belief degrees of referential values and support values of sensor data. Belief transaction database and support values of sensor data are used as input for this procedure. This procedure results cell values of belief matrix obtained by multiplying corresponding belief degrees and support values of the attributes in a sensor data as can be seen in expression (18), which is used as input for confidence calculation. Since sensor data value, which is quantitative in nature, is distributed over different referential values to address semantic poorness of the sensor data as shown in Tables 4 and 5. The belief degrees attached to referential values corresponding to the sensor data address the uncertainty due to ambiguity, imprecision and vagueness. However, this is unable to remove uncertainty due to incompleteness, and hence, the belief degrees associated with referential values are required to be multiplied with corresponding support values of the sensor data to remove the uncertainty due to incompleteness (Yang et al. 2006). In this way, a belief matrix can be formed by using the expression (18) and elaborated in Table 7.

In expression (18), $Belief\_Matrix\_Element_{i,j}$ represents each element of belief matrix, $sup(a_k)$ represents the support value of sensor data a_{k_i} for attribute a_k, and $Belief\_Tran\_Database_{k_i,x}$ represents a referential value of sensor data a_{k_i}.

$$Belief\_Matrix\_Element_{i,j} = sup(a_{k_i}) \times Belief\_Tran\_Database_{k_i,x} \quad (18)$$

The belief transaction database of rainfall and temperature is transformed in to belief matrix which is shown in Table 7. Columns 1 to 10 of Table 7 shows the values of belief matrix computed by using expression (18). As an example, the cell (1, 1) of belief matrix, which is 0.6902 is obtained by applying expression (18).

3.3.5 Confidence calculation

Confidence is an assessment of the degree of certainty of the identified association between antecedent and consequent of a rule. Rule activation, as shown in expression (10), of belief rule-based inference mechanism is quite similar to confidence calculation. However, combined matching degree, as shown in expression (11), of rule activation is calculated using multiplicative aggregation function. Since it is not suitable for sensor data due to its nature, which can be replaced by a popular similarity measure named hamming distance (Hamming 1950). In addition, hamming distance is suitable to work with sensor data, as it is particularity designed to work with quantitative data, which is a common feature of sensor data. Therefore, confidence of the belief-rule-based association rule can be defined as a function of hamming distance (Black 2004; Hamming 1950) of the transactions and total summation of hamming distance of all the transactions of the belief matrix.

The expression (19) calculates the hamming distance for a transaction in respect of other transactions, and then summation of all the distances is assigned in $\alpha_{t_i,k}$ for transaction

Table 7 Belief matrix for rainfall and temperature

t1	t2	t3	t4	t5	t6	t7	t8	t9	t10
0.6902	0.0	4.5	12.18	11.78	0.0	0.0	10.5398	12.4838	7.055
0.0098	0.0	40.5	16.82	7.22	49.98	48.9918	24.3602	13.4162	1.445
0.0	0.0	0.0	0.0	0.0	1.02	8.1082	0.0	0.0	0.0
0.0	38.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	152.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
0.81	0.0	0.0	0.0	0.0	0.0	0.81	0.0	0.0	1.44
1.89	2.56	0.0	2.56	3.0	3.0	1.89	2.79	3.0	0.96
0.0	0.64	4.0	0.64	0.0	0.0	0.0	0.31	0.0	0.0

Table 8 Final result

Transaction ID	Confidence value	Rainfall	Temperature
t1	0.08	7	27
t2	0.011	1900	32
t3	0.08	450	40
t4	0.07	290	32
t5	0.11	190	30
t6	0.13	510	30
t7	0.08	571	27
t8	0.07	349	31
t9	0.08	259	30
t10	0.19	85	24

t_i and attribute k. The expression (20) sums all the $\alpha_{t_i,k}$ and assigns to θ_k. Finally, the confidence of each transaction can be obtained by using the expression (21).

$$\alpha_{t_i,k} = \sum_{j=1}^{n} Hamming\_Distance(t_j) \tag{19}$$

$$\theta_k = \sum_{i=1}^{n} \alpha_{t_i,k} \tag{20}$$

$$Confidence(t_i) = \frac{\sum_{k=1}^{m} \alpha_{t_i,k}}{\sum_{k=1}^{m} \theta_k} \tag{21}$$

As for example, $\alpha_{t_i,k}$ (where $i = 1$ and $k = rainfall$) is $(0+4+0+0+0+2+2+0+0+0) = 8$ by using expression (19). Likewise, $\alpha_{t_i,k}$(where $i = 1$ and $k = temperature$) is $(0+2+3+2+1+1+0+2+1+0) = 12$ by using expression (19). θ_k (where $k = rainfall$) is $(8+36+8+8+8+18+18+8+8+8) = 128$ and θ_k (where $k = temperature$) is $(12+10+18+10+8+8+12+10+8+12) = 108$ by using expression (20). Therefore, $Confidence(t_1)$ is $(8+12)/(128+108) = 0.08$ by using expression (21). Table 8 shows the confidence value for each of the transactions. Confidence values for each of the transaction are shown in Column 2 of Table 8. Sensor data of rainfall and temperature are shown in Columns 3 and 4 of Table 8, respectively.

3.3.6 Belief association rule discovery

Belief association rule discovery procedure consists of BRBAR based on referential values from belief transaction database and confidence values discovered in the previous procedure. Traditional belief rules, as shown in expression (8), consist of belief degrees in consequent part of the rule due to unavailability of the belief degree for the referential values of antecedents. However, the belief degrees are embedded with the referential values of antecedent and consequent part of new belief association rule [expression (22)],which can be discovered from the belief transaction database. This makes the belief association rule more robust than the belief rules. Therefore, a novel belief rule named belief association rule is proposed here.

In consultation with experts, a threshold value is selected for confidence value, which filter outs anomalous transaction from belief transaction database. Subsequently, Belief association rule is created by embedding belief degrees associated with the referential values in the antecedent and consequent part of a rule from the belief transaction database. Eventually, these belief association rules will be used as initial rule base for belief-rule-based expert system.

As an example, expression (22) represents a belief association rule, which can be interpreted as rainfall with 98.6% probability of no rainfall and 1.4% probability of low implies temperature of 30% probability of medium and 70% probability of hot.

$$R_1: \begin{cases} Rainfall\{(NoRainfall, 0.986)(Low, 0.014)(Medium, 0.0)(High, 0.0) \\ \qquad (VeryHigh, 0.0)\} \\ \Longrightarrow Temp\{(VeryLow, 0.0)(Low, 0.0)(Medium, 0.3)(Hot, 0.7) \\ \qquad (VeryHot, 0.0)\} \end{cases} \tag{22}$$

$$R_2: \begin{cases} Rainfall\{(NoRainfall, 0.0)(Low, 0.0)(Medium, 0.0)(High, 0.2) \\ \qquad (VeryHigh, 0.8)\} \\ \Longrightarrow Temp\{(VeryLow, 0.0)(Low, 0.0)(Medium, 0.0)(Hot, 0.8) \\ \qquad (VeryHot, 0.2)\} \end{cases} \tag{23}$$

In summary, a novel anomaly detection procedure, named BRBAR is proposed for detecting anomaly from sensor data. The new BRBAR is able to address different types of uncertainty like incompleteness, ignorance, vagueness, imprecision and ambiguity, which are common features of sensor data. A new support calculation procedure is proposed, which addresses uncertainties due to incompleteness. Furthermore, an improved and sensor data friendly confidence calculation method is proposed by using hamming distance instead of using multiplicative aggregation function. Moreover, a robust belief association rule is proposed by embedding belief degrees with the referential values in antecedent part of the rule, which will be used as initial rule base for expert system. Henceforth, in the next section anomaly-free sensor data will be fed into a belief-rule-based expert system to show the effects of the new BRBAR.

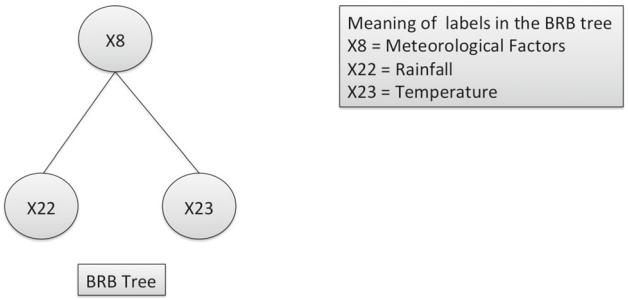

Fig. 2 The belief-rule-based tree

4 Feeding nonanomalous data into belief-rule-based expert system

A web-based belief-rule-based expert system (Web-BRBES) (Islam et al. 2015) is capable of handling sensor data as well as enabling flood prediction, and this system is fed with the rainfall and temperature sample sensor data considered in this research as shown in Table 1. A portion of the flood prediction BRB tree as mentioned in Web-BRBES (Islam et al. 2015) is considered for the demonstration of BRBAR algorithm, as shown in Fig. 2. The root node of this tree (X8) represents "Metrological Condition", and two leaf nodes X22 and X23 represent "Rainfall" and "Temperature", respectively.

Figure 3 shows the input and output of the Web-BRBES for anomalous data. The right and left square boxes in Fig. 3 show the graph of input and output data of Web-BRBES named input? and output?, respectively. In the input graph, the X-axis displays the data in chronological order, while the Y-axis displays values of the data gathered by sensors during each time interval. In a similar manner, the output graph of Fig. 3 can be explained. Figure 4 shows the input and output for anomaly-free data, which can be understood in similar way.

In the input graph of Fig. 3, an unusual peak can be seen on Y-axis for X22 (rainfall), which influences the output of Web-BRBES with two peaks in the output graph. Therefore, removing the anomalous sensor data by using BRBAR a different output values is seen in output graph of Fig. 4. Moreover, the average crisp value of "Metrological Condition" for anomalous sensor data is 0.03255 and for anomaly-free sensor data is 0.02105. This shows that due to anomalous data appropriate value for "Metrological Condition" cannot be found, which in turns hamper the prediction of flood water level.

5 Performance evaluation of the belief-rule-based expert system

The comparison, evaluation and assessment of the accuracy of the results generated from the different models or techniques are considered as an important aspect to measure the reliability of a research. Receiver operator characteristic (ROC) curves are widely used to evaluate, compare and assess the performance of different methods and techniques. The reason for this is that it provides a comprehensive and visual methods of summarising the

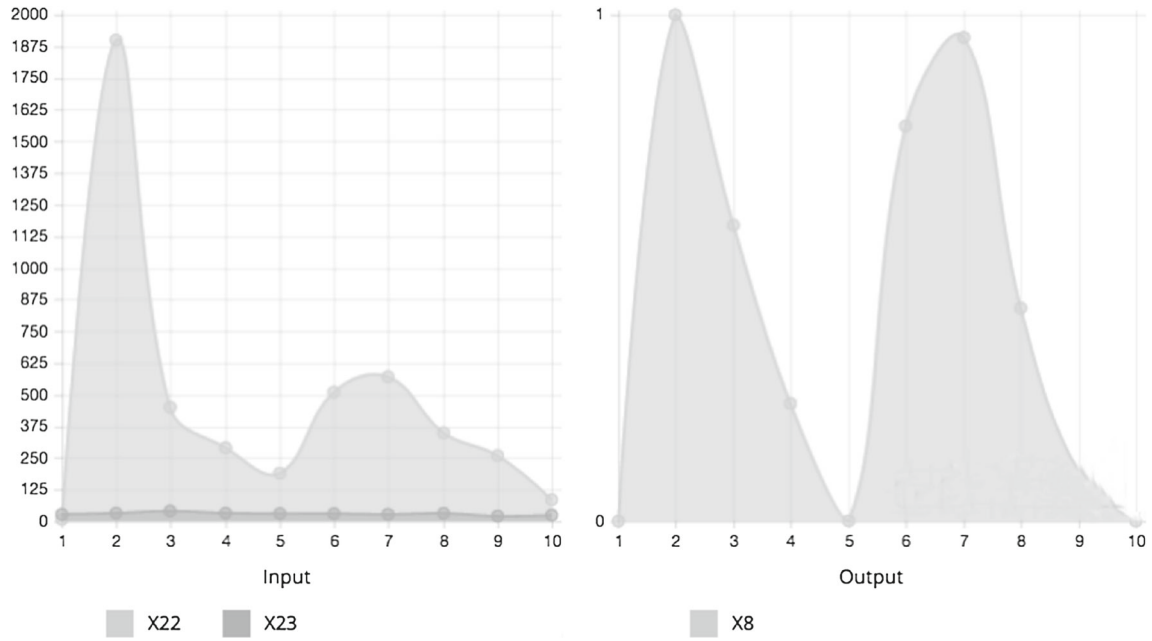

Fig. 3 Output of Web-BRBES for sensor data with anomaly

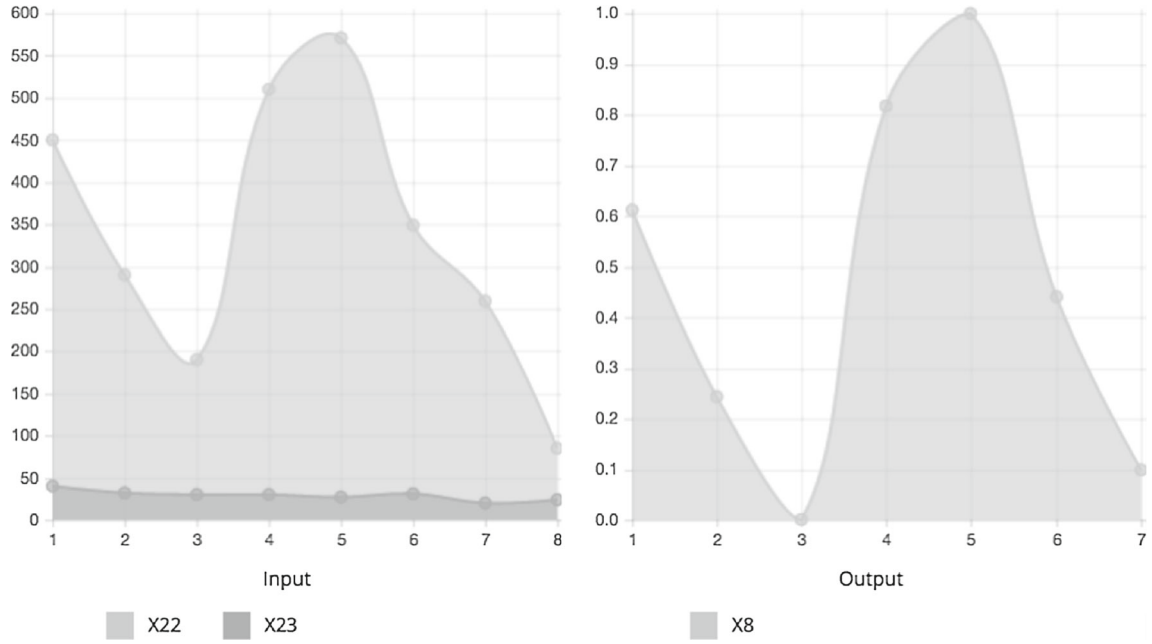

Fig. 4 Output of Web-BRBES for sensor data without anomaly

accuracy of comparison, evaluation and assessment (Hossain et al. 2015c; Gönen 2007). Thus, ROC curves have become the prominent tool for evaluating different models, algorithms and techniques in various research fields such as machine learning, clinical applications, atmospheric science and many others (Zou et al. 2007; Karim et al. 2016; Hossain et al. 2015a). Therefore, in this research ROC curves were used to measure the accuracy of anomaly detection using BRBAR and compare its performance with other similar techniques such as, Gaussian-based anomaly detection, binary and fuzzy association rules. In ROC curves, the accuracy can be measured by calculating the size of the Area under curve (AUC) (Gagnon and Peterson 1998). The larger the area, the higher is the accuracy of the results.

To evaluate the performances of BRBAR by using ROC curves, rainfall and temperature sensor data collected from Climate Division of Bangladesh Meteorological Department (2016) have been considered. In addition to the sensor data, to investigate the applicability of the developed novel anomaly detection algorithm in other domains Breast Cancer Wisconsin dataset collected from the UCI machine learning repository has also been considered.

Experts' perception on the anomaly and nonanomaly of sensor data has been considered as the baseline to do the comparison among BRBAR, Gaussian, binary and fuzzy association rules. However, for breast cancer data appropriate diagnostic result investigation of the disease has been considered as the baseline. The rainfall and temperature sensor dataset consists of 380 readings of sensor data of Chittagong in Bangladesh. The rainfall is measured in millimetre and temperature in Celsius. These sample data of rainfall and temperature can be considered sufficient, because sample sizes of more than 30 and less than 500 are appropriate for most research (Roscoe 1975). The dataset for Breast Cancer Wisconsin (Diagnostic) consists of 669 records with 8 attributes. This dataset contains different characteristics of cancer cells. Furthermore, this dataset is labelled (benign or malignant) with the status of the cancer cell. Although the breast cancer dataset is more than 500 records, the 669 records which used in this research are considered as standard (Karim et al. 2016; Hossain et al. 2016).

Figure 5 shows ROC curves for Gaussian, binary association rule, fuzzy association rule and BRBAR for the rainfall and temperature data. The AUC and confidence interval (CI) for above techniques are shown in Table 9. The area under curve for Gaussian, binary association rule, fuzzy association rule and BRBAR are 0.168, 0.843, 0.867 and 0.990, respectively, as shown in Table 9. It can be observed from the results shown in Table 9 that the coverage of BRBAR is better than the other mentioned techniques. This implies anomaly detection from sensor data by BRBAR has performed better than the other techniques due to addressing of different types of uncertainty like incompleteness, ignorance, vagueness, imprecision and ambiguity.

Figure 6 shows the ROC curves for the above-mentioned techniques. The area under curve for Gaussian, binary association rule, fuzzy association rule and BRBAR is 0.472,

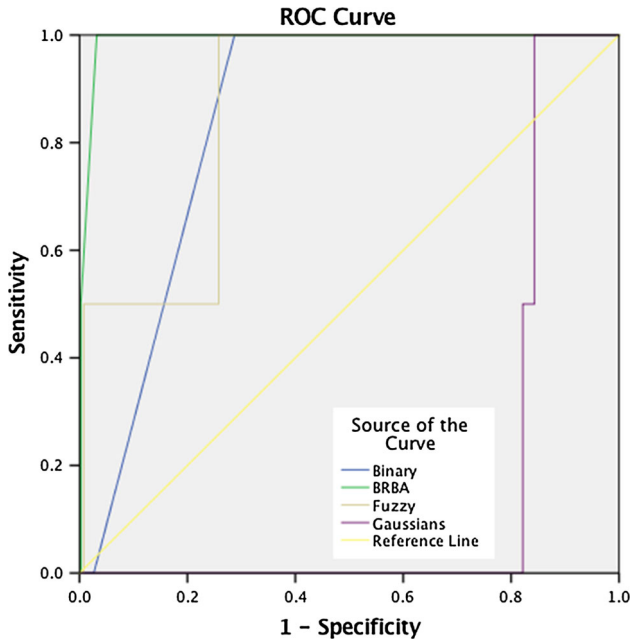

Fig. 5 ROC curve comparison of binary, fuzzy, BRB association rule and Gaussian for rainfall and temperature data

Table 9 Comparison of different techniques for rainfall and temperature. CI 95% confidence interval

Results	Gaussian	Binary association rule	Fuzzy association rule	BRBAR
AUC	0.168	0.843	0.867	0.990
CI	0.127–0.208	0.734–0.953	0.692–1.0	0.974–1.0

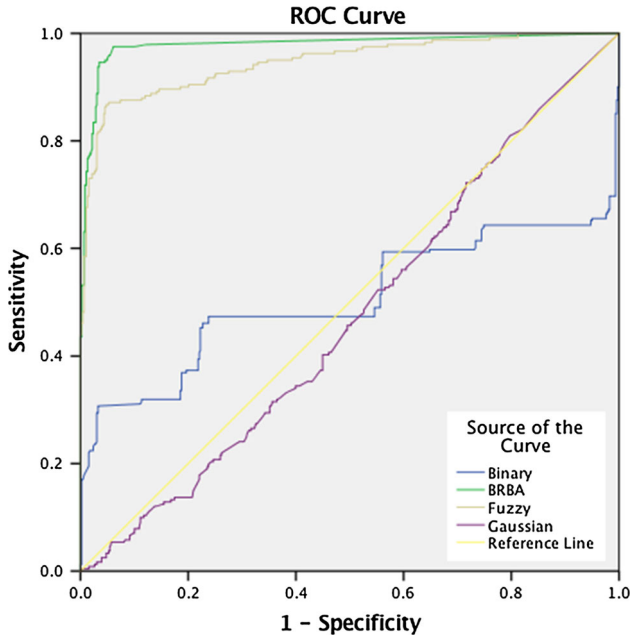

Fig. 6 ROC curve comparison of binary, fuzzy, BRB association rule and Gaussian for Breast Cancer Wisconsin (diagnostic) data

Table 10 Comparison of different techniques for Breast Cancer Wisconsin (diagnostic) data. CI 95% confidence interval

Results	Gaussian	Binary association rule	Fuzzy association rule	BRBAR
AUC	0.472	0.505	0.946	0.979
CI	0.428–0.516	0.450–0.560	0.927–0.965	0.967–0.991

0.505, 0.946 and 0.979, respectively, as shown in Table 10. From the above results, it can be observed that the coverage of BRBAR is better than the other mentioned techniques. This also shows that BRBAR performs well for anomaly detection in normal data.

It is evident (Tables 9, 10; Figs. 5, 6) that Gaussian anomaly detection technique performed comparatively better for breast cancer data than from rainfall and temperature data. Gaussian anomaly detection technique assumes that the sample data follow normal distribution (see Sect. 2). Therefore, it performs poorly for rainfall and temperature data as the sensor data do not follow normal distribution. Moreover, Gaussian does not address any types of uncertainty, which also influences the performance of anomaly detection. On the contrary, association rule does not depend on the distribution of the data. Therefore, it performs better than the Gaussian algorithms for the both datasets. However, due to lack of addressing any types of uncertainty binary association rule does not perform better than the fuzzy and BRBAR (see Sect. 3.1). Fuzzy association rule handles uncertainties due to imprecision, ambiguity and vagueness which helps it to perform better then the binary association rule (see Sect. 3.2). Therefore, by addressing uncertainties due to imprecision, ambiguity and vagueness fuzzy association rule performs better than the binary association rule. Finally, BRBAR address all types of uncertainty in an integrated framework, which leads to the better performance than from the Gaussian, binary and fuzzy association rules. From the above discussion, it can be observed that anomaly detection from sensor data by BRBAR performs better than the other techniques due to addressing of different types of uncertainty like incompleteness, ignorance, vagueness, imprecision and ambiguity. In addition, BRBAR performs better not only in anomaly detection for sensor data, but also for other domains such as breast cancer. Moreover, BRBAR does not depend on any training dataset for anomaly detection like supervised and semi-supervised machine learning algorithms (Chandola et al. 2009). Therefore, BRBAR will outperform the above-mentioned algorithms as the accuracy of them depend on the training of the supervised and semi-supervised machine learning algorithms.

6 Conclusion

A novel anomaly detection algorithm for sensor data based on BRBAR is proposed in this research work. The BRBAR has the capability of handling different kinds of uncertainty such as incompleteness, ignorance, vagueness, imprecision and ambiguity, which are common features of sensor data (see Sect. 3.3). Due to the nature of sensor data, the traditional inference mechanism of belief rule cannot be used. Therefore, a new inference mechanism is proposed, which consists of input transaction database, converting into belief transaction database, support calculation, belief matrix, confidence calculation and belief association rule discovery.

A new support calculation procedure is proposed, which addresses uncertainties due to incompleteness (see Sect. 3.3.3). In addition, an improved and sensor data friendly confidence calculation method is proposed by using hamming distance instead of using multiplicative aggregation function (see Sect. 3.3.5). Since, hamming distance is more suitable for sensor data then multiplicative aggregation function as sensor data is more quantitative in nature (Calzada et al. 2014). Moreover, a robust belief association rule is proposed by embedding belief degrees with the referential values in antecedent part of the rule, which will be used as initial rule base for expert system (see Sect. 3.3.6). Since, traditional belief rule lacks belief degrees with respect to the referential values of antecedent part of the rule [expression (22)]. The results of BRBAR have been compared against three other anomaly detection techniques (such as, Gaussian, binary association rule and fuzzy association rule) with two different types of datasets. It has been demonstrated that BRBAR performed better than the other techniques for both the datasets (see Figs. 5, 6). The reason for this is Gaussian, is a statistical-based approach, and is unable to handle uncertainty due to incompleteness, ignorance, vagueness, imprecision and ambiguity, while binary association rule uses assertive knowledge, which can be evaluated either true or false. Hence, this approach is unable to address any type of uncertainty. On the contrary, fuzzy association rule can handle uncertainty due to vagueness, ambiguity and imprecision but unable to handle ignorance and incompleteness. However, BRBAR can handle all types of uncertainty in an integrated framework. Moreover, both Gaussian and binary association rule lack the better representation of semantic content, and hence, uncertainty due to linguistic labels cannot be addressed by using these methods. In addition, the ROC curves (see Figs. 5, 6) show that AUC of BRBAR is better than the above-mentioned techniques, because the proposed technique in this paper handles all types of uncertainty as mentioned. The proposed anomaly detection algorithm demonstrates a way of extracting initial belief rule base from sensor data, which can be considered as a significant contribution in the area of knowledge acquisition.

Moreover, anomaly-free sensor data as well as anomalous sensor data are fed into the Web-BRBES. It can be observed that Web-BRBES provides better result of detecting metrological condition for anomaly-free data than from the anomalous data (see Figs. 3, 4; Sect. 4). In addition, BRBAR helps Web-BRBES to perform more reliable and accurate prediction of flood, using the data received from sensors, deployed in a flood prone area by removing the anomalies. Hence, it can be argued that the novel BRBAR technique will improve anomaly detection approach for other application areas such as, surveillance, environmental monitoring and disaster management under uncertainty. This new anomaly detection algorithm will also improve the prediction of different expert systems as anomalous data can be removed more efficiently.

In this research work, preliminarily BRBAR has been tested with two datasets. However, the performance of the algorithm needs to be tested by using more data from different types of sensor to ensure its efficiency and robustness. In addition, more investigation is needed for choosing appropriate benchmark data. Furthermore, as a future work, more research can be carried out for BRB inference mechanism for initial rule base coming out from BRBAR. In addition, investigation on benchmark data and testing the algorithm with different sensor data can also be considered as future work.

Funding This study was funded by Swedish Research Council under Grant 2014-4251.

Compliance with ethical standards

Conflict of interest Authors Raihan Ul Islam, Mohammad Shahadat Hossain, and Karl Andersson declare that they have no conflict of interest.

Human and animal rights This article does not contain any studies with human participants or animals performed by any of the authors.

References

Adefisan EA, Bayo AS, Ropo OI (2015) Application of geo-spatial technology in identifying areas vulnerable to flooding in ibadan metropolis. J Environ Earth Sci 5(14):153–166

Agrawal R, Srikant R (1994) Fast algorithms for mining association rules. In: Proceedings of the 20th international conference very large data bases, VLDB, vol 1215, pp 487–499

Ahmad N, Hussain M, Riaz N, Subhani F, Haider S, Alamgir KS, Shinwari F (2013) Flood prediction and disaster risk analysis using gis based wireless sensor networks, a review. J Basic Appl Sci Res 3(8):632–643

Andersson K, Hossain MS (2014) Smart risk assessment systems using belief-rule-based DSS and WSN technologies. In: Proceedings of 2014 4th international conference on wireless communications, vehicular technology, information theory and aerospace electronic systems (VITAE), pp 1–5

Andersson K, Hossain MS (2015) Heterogeneous wireless sensor networks for flood prediction decision support systems. In: Proceedings of the 2015 IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp 133–137

Atzori L, Iera A, Morabito G (2010) The Internet of Things: a survey. Comput Netw 54(15):2787–2805

Aziz NAA, Aziz KA (2011) Managing disaster with wireless sensor networks. In: 2011 13th international conference on advanced communication technology (ICACT), pp 202–207. IEEE

Bajaber F, Awan I (2010) Energy efficient clustering protocol to enhance lifetime of wireless sensor network. J Ambient Intell Humaniz Comput 1(4):239–248

Barnett V, Lewis T (1994) Outliers in statistical data. 3rd edn. John Wiley Sons, New York

Barnet V (1976) The ordering of multivariate data (with discussion). J R Stat Soc Ser A 139:318–354

Beckman RJ, Cook RD (1983) Outlier. s. Technometrics 25(2):119–149

Black PE (2004) Dictionary of algorithms and data structures. National Institute of Standards and Technology, Gaithersburg

Calzada A, Liu J, Nugent CD, Wang H, Martinez L (2014) Sensor-based activity recognition using extended belief rule-based inference methodology. In: 2014 36th annual international conference of the IEEE engineering in medicine and biology society, pp 2694–2697. IEEE

Chandola V, Banerjee A, Kumar V (2009) Anomaly detection: a survey. ACM Comput Surv (CSUR) 41(3):15

Chen Z, Chen G (2007) An approach to classification based on fuzzy association rules. In: Proceedings of the international conference on intelligent systems and knowledge engineering

Chen J, Kher S, Somani A (2006) Distributed fault detection of wireless sensor networks. In: Proceedings of the 2006 workshop on dependability issues in wireless ad hoc networks and sensor networks

Climate division of bangladesh meteorological department (2016). http://www.bmd.gov.bd/?/home/[15.10.2016]

Demeritt D, Nobert S, Cloke HL, Pappenberger F (2013) The European Flood Alert System and the communication, perception, and use of ensemble predictions for operational flood risk management. Hydrol Process 27(1):147–157

Dhanya CT, Kumar DN (2009) Data mining for evolving fuzzy association rules for predicting monsoon rainfall of india. J Intell Syst 18(3):193–210

Fang S, Xu LD, Zhu Y, Ahati J, Pei H, Yan J, Liu Z (2014) An integrated system for regional environmental monitoring and management based on Internet of Things. IEEE Trans Ind Inform 10(2):1596–1605

Fiore U, Palmieri F, Castiglione A, De Santis A (2013) Network anomaly detection with the restricted boltzmann machine. Neurocomputing 122:13–23

Gagnon RC, Peterson JJ (1998) Estimation of confidence intervals for area under the curve from destructively obtained pharmacokinetic data. J Pharmacokinet Biopharm 26(1):87–102

Gnecco G, Morisi R, Roth G, Sanguineti M, Taramasso AC (2016) Supervised and semi-supervised classifiers for the detection of flood-prone areas. Soft Comput 1–13

Gönen M (2007) Analyzing receiver operating characteristic curves with SAS. SAS Institute, Cary NC

González A, Donnelly A, Jones M, Chrysoulakis N, Lopes M (2013) A decision-support system for sustainable urban metabolism in Europe. Environ Impact Assess Rev 38:109–119

Hamming RW (1950) Error detecting and error correcting codes. Bell Syst Tech J 29(2):147–160

He Z, Xu X, Huang JZ, Deng S (2004) A frequent pattern discovery method for outlier detection. In: Advances in web-age information management. Springer, LNCS 3129, pp 726–732

Hossain MS, Davies CG (2001) A system for the simulation and prediction of floods. In: Proceedings of the GIS research UK (GISRUK 2001)

Hossain MS, Davies CG (2004) A GIS to reduce flood impact on road transportation systems. In: Proceedings of 2004 ESRI user conference

Hossain MS, Davies CG (2006) An information system to assess flood impact on road transportation systems. Int J Comput Appl 13(2):73–80

Hossain MS, Hossain ME, Khalid MS, Haque MA (2014) A belief rule-based (BRB) decision support system for assessing clinical asthma suspicion. In: Proceedings of the scandinavian conference on health informatics, pp 83–89

Hossain MS, Hasan MA, Uddin M, Islam MM, Mustafa R (2015a) A belief rule based expert system to assess lung cancer under uncertainty. In: 2015 18th international conference on computer and information technology (ICCIT), pp 413–418. IEEE

Hossain MS, Andersson K, Naznin S (2015b) A belief rule based expert system to diagnose measles under uncertainty. In: Proceedings of the 2015 international conference on health informatics and medical systems (HIMS'15), pp 17–23

Hossain MS, Zander P, Kamal MS, Chowdhury L (2015c) Belief-rule-based expert systems for evaluation of e-government: a case study. Expert Syst 32(5):563–577

Hossain MS, Haque MA, Mustafa R, Karim R, Dey HR, Yousuf M (2016) An expert system to assist the diagnosis of ischemic heart disease. Int J Integr Care

Islam R Ul, Andersson K, Hossain MS (2015) A web based belief rule based expert system to predict flood. In: Proceedings of the 17th International conference on information integration and web-based applications & services, pp 19–26. ACM

Karim R, Hossain MS, Andersson K, Uddin MJ, Meah MP (2016) A belief rule based expert system to assess clinical bronchopneumonia suspicion. In: Proceedings of future technologies conference 2016 (FTC 2016)

Khedo K (2013) Real-time flood monitoring using wireless sensor networks. J Inst Eng Maurit (IEM), Special Issue: Disaster and Risk Management, pp 59–69

Langin C, Rahimi S (2010) Soft computing in intrusion detection: the state of the art. J Ambient Intell Humaniz Comput 1(2):133–145

Martí L, Sanchez-Pi N, Manuel Molina J, Garcia ACB (2015) Anomaly detection based on sensor data in petroleum industry applications. Sensors 15(2):2774–2797

Muyeba M, Khan MS, Coenen F (2008) Fuzzy weighted association rule mining with weighted support and confidence framework. In: Proceedings of Pacific-Asia conference on knowledge discovery and data mining, pp 49–61

Palattella MR, Accettura N, Vilajosana X, Watteyne T, Grieco LA, Boggia G, Dohler M (2013) Standardized protocol stack for the Internet of (Important) Things. IEEE Commun Surv Tutor 15(3):1389–1406

Pappenberger F, Matgen P, Beven KJ, Henry JB, Pfister L, Fraipont P (2006) Influence of uncertain boundary conditions and model structure on flood inundation predictions. Adv Water Resour 29(10):1430–1449

Patcha A, Park JM (2007) An overview of anomaly detection techniques: existing solutions and latest technological trends. Comput Netw 51(12):3448–3470

Perrig A, Stankovic J, Wagner D (2004) Security in wireless sensor networks. Commun ACM 47(6):53–57

Rahaman S, Hossain MS (2013) A belief rule based clinical decision support system to assess suspicion of heart failure from signs, symptoms and risk factors. In: 2013 international conference on informatics, electronics & vision (ICIEV), pp 1–6. IEEE

Rajasegarar S, Leckie C, Palaniswami M (2008) Anomaly detection in wireless sensor networks. IEEE Wirel Commun 15(4):34–40

Rajeswari, AM, Sridevi M, Deisy C (2014) Outliers detection on educational data using fuzzy association rule mining. In: Proceedings of International Conference on Advanced in Computer Communication and Information Science (ACCIS-14), pp 1–9

Roscoe JT (1975) Fundamental research statistics for the behavioural sciences. Rinehart and Winston, New York

Ruiz MD, Martin-Bautista MJ, Sánchez D, Vila MA, Delgado M (2014) Anomaly detection using fuzzy association rules. Int J Electron Secur Digit Forensics 96(1):25–37

Ruiz M, Martin-Bautista M, Sanchez D, Delgado M (2016) Discovering fuzzy exception and anomalous rules. IEEE Trans Fuzzy Syst 24(4):930–944

Seal V, Raha A, Maity S, Mitra SK, Mukherjee A, Naskar MK (2012) A simple flood forecasting scheme using wireless sensor networks. Int J Ad Hoc Sens Ubiquitous Comput (IJASUC) 3(1):45–60

Sheltami TR, Bala A, Shakshuki EM (2016) Wireless sensor networks for leak detection in pipelines: a survey. J Ambient Intell Humaniz Comput 7(3):347–356

Thombre S, Islam R Ul, Andersson K, Hossain MS (2016) Performance analysis of an IP based protocol stack for WSNs. In: Proceedings of the 2016 IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp 691–696

Vladimirova T, Yuhaniz S (2011) An intelligent decision-making system for flood monitoring from space. Soft Comput 15(1):13–24

Wang YM, Yang JB, Xu DL (2006) Environmental impact assessment using the evidential reasoning approach. Eur J Oper Res 174(3):1885–1913

Weng CH (2011) Mining fuzzy specific rare itemsets for education data. Knowl Based Syst 24(5):697–708

Wijsen J, Meersman R (1998) On the complexity of mining quantitative association rules. Data Min Knowl Discov 2(3):263–281

Xiea Q, Sua Z, Zhenga P, Liuc J, Yud H (2014) Multi-sensor data fusion based on fuzzy neural network and its application in piggery environmental control strategies. J Inf Comput Sci 11(15):5407–5418

Xu DL, Liu J, Yang JB, Liu GP, Wang J, Jenkinson I, Ren J (2007) Inference and learning methodology of belief-rule-based expert system for pipeline leak detection. Expert Syst Appl 32(1):103–113

Xu L, Zhang J, Tsai PW., Wu W, Wang DJ (2015) Uncertain random spectra: a new metric for assessing the survivability of mobile wireless sensor networks. Soft Comput 1–11. doi:10.1007/s00500-015-1962-4

Yang JB, Sen P (1994) A general multi-level evaluation process for hybrid MADM with uncertainty. IEEE Trans Syst Man Cybern 24(10):1458–1473

Yang JB, Liu J, Wang J, Sii HS, Wang HW (2006) Belief rule-base inference methodology using the evidential reasoning approach-RIMER. IEEE Trans Syst Man Cybern Part A Syst Hum 36(2):266–285

Yuan Y, Feldhamer S, Gafni A, Fyfe F, Ludwin D (2002) The development and evaluation of a fuzzy logic expert system for renal transplantation assignment: Is this a useful tool? Eur J Oper Res 142(1):152–173

Zhang Y, Meratnia N, Havinga P (2010) Outlier detection techniques for wireless sensor networks: a survey. IEEE Commun Surv Tutor 12(2):159–170

Zhang J, Gao Q, Wang H, Wang H (2011) Detecting anomalies from high-dimensional wireless network data streams: a case study. Soft Comput 15(6):1195–1215

Zou KH, O'Malley AJ, Mauri L (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. Circulation 115(5):654–657